

Testing Lipschitz Property over Product Distribution and its Applications to Statistical Data Privacy

Kashyap Dixit
Pennsylvania State University
kashyap@cse.psu.edu

Madhav Jha
Pennsylvania State University
mxj201@cse.psu.edu

Abhradeep Thakurta
Pennsylvania State University
azg161@cse.psu.edu

Abstract

Analysis of statistical data privacy has emerged as an important area of research. In this work we design algorithms to test privacy guarantees of a given Algorithm \mathcal{A} executing on a data set \mathcal{D} which contains potentially sensitive information about individuals. We design an efficient algorithm \mathcal{A}_{test} which can verify whether \mathcal{A} satisfies *generalized differential privacy* guarantee. Generalized differential privacy [BBG⁺11] is a relaxation of the notion of differential privacy initially proposed by [DMNS06]. By now differential privacy is the most widely accepted notion of statistical data privacy.

To design Algorithm \mathcal{A}_{test} , we show a new connection between the differential privacy guarantee and Lipschitzness property of a given function. More specifically, we show that an efficient algorithm for testing of Lipschitz property can be transformed into \mathcal{A}_{test} which can test for generalized differential privacy. Lipschitz property testing and its variants, first studied by [JR11], has been explored by many works [JR11, AJMR12b, AJMR12a, CS12] because of its intrinsic connection to data privacy as highlighted by [JR11]. To develop a Lipschitz property tester with an explicit application in privacy has been an intriguing problem since the work of [JR11]. In our work, we present such a direct application of lipschitz tester to testing privacy. We provide concrete instantiations of Lipschitz testers (over both the hypercube and the hypergrid domains) which are used in \mathcal{A}_{test} to test for privacy of Algorithm \mathcal{A} when the underlying data set \mathcal{D} is drawn from the hypercube and the hypergrid domains respectively.

Apart from showing a direct connection between testing of privacy and Lipschitzness testing, we generalize the work of [JR11] to the setting of distribution property testing. We design an efficient Lipschitz testing algorithm when the distribution over the domain points is not uniform. More precisely, we design an efficient Lipschitz tester for the case where the domain points are drawn from hypercube according to some fixed product distribution. This result is of independent interest to the property testing community. It is important to note that to the best of our knowledge our results on Lipschitz testing over product distributions is the only positive result in property testing literature for non-uniform distributions after [AC06].

1 Introduction

Consider a data sharing platform like *BlueKai*, *TellApart* or *Criteo*. These platforms extensively collect and share user data with third-parties (e.g., advertisers) to enhance specific user experience (e.g., better behavioral targeting). Now, the third party applications use these data to train their machine learning algorithms for better prediction abilities. Since, the data which gets shared is extremely rich in user information, it immediately poses privacy concerns over the user information [Kor10, CKN⁺11]. One way to address the privacy concerns due to the third-party learning algorithms is to train the third party algorithms “in-house”, i.e., within the data sharing platform itself thus, making sure that the trained machine learning model preserves privacy of the underlying training data. In this paper, we study a theoretical abstraction of the above mentioned problem.

Let \mathcal{D} be a data set where each record corresponds to a particular user and contains potentially sensitive information about the user (for example, the click history of the user for a set of advertisements displayed). Let \mathcal{A} be an algorithm that we would like to execute on the data set \mathcal{D} (possibly to obtain some global trends about the users in \mathcal{D}) without compromising individual’s privacy. This challenging problem has recently received a lot of attention in the form of theoretical investigation in determining the privacy-utility trade-offs for various old and new algorithms. However, even if an algorithm is provably “safe”, in practice the algorithm will be implemented in a programming language that may originate from untrusted third party. This brings its own set of challenges and has primarily been addressed in the following way: transform the algorithm \mathcal{A} into a variant which provably satisfies some theoretically sound notion of data privacy (e.g., *differential privacy* [DMNS06]) either by syntactic manipulation (e.g. [McS09, RP10]) or doing so in some algorithmic/systems framework (eg. [NRS07, JR11, MTS⁺12, RSK⁺10]). While each approach has its own appeal, they all have a few shortcomings. For example, they suffer from weak utility guarantees [NRS07, MTS⁺12, RSK⁺10] or take prohibitively large running time [JR11] or require use of specialized syntax [McS09, RP10] making it somewhat nontrivial for a non-privacy expert to produce an effective transformation.

In this work, we take a new approach to the above problem which we call *privacy testing*. Specifically, we initiate the study of *testing* whether an input algorithm \mathcal{A} satisfies statistical privacy guarantees. We do this by formulating the problem in the well-studied framework of property testing [RS96a, GGR98a].

Privacy testing Before we execute an Algorithm \mathcal{A} which claims to satisfy a pre-approved notion of privacy, we test for the validity of such a claim. To the best of our knowledge, ours is the first work to study this approach. More precisely, in this work we initiate the study of testing an algorithm \mathcal{A} for differential privacy guarantees. Differential privacy in the recent past has become a well established notion of privacy [Dwo06, Dwo08, Dwo09]. Roughly speaking, differential privacy guarantees that the output of an algorithm \mathcal{A} will not depend “too much” on any particular record of the underlying data set \mathcal{D} . We design testing algorithms to test whether \mathcal{A} satisfies *generalized* differential privacy [BBG⁺11] or not. *Generalized differential privacy* is a relaxation of differential privacy and follows the same principles as differential privacy. Under specific setting of parameters, generalized differential privacy collapses to the definition of differential privacy. For a precise definition, see Section 2.1. It seems to us (and we make it more formal later on) that it may not be possible to design a computationally efficient testing algorithm for testing the notion of exact differential privacy, since in some sense it is a worst case notion privacy (see [BBG⁺11, BD12] for a discussion on this).

Testing Lipschitz property under product distribution and its connection to privacy testing The goal of testing properties of functions is to distinguish between functions which satisfy a given property from functions which are “far” from satisfying the property. The notion of “far” is usually the fraction of points in the domain of the function on which the function needs to be redefined to make it satisfy the property.

To test for generalized differential privacy, we show a new connection between differential privacy and the problem of testing Lipschitz property which was first studied by [JR11]. A recent line of work [JR11, AJMR12b, AJMR12a] has sought to explore applications of sublinear algorithms (specifically, *property testers* and *reconstructors*) to data privacy. We continue this line of work and show the first application of property testers (which

are vastly more efficient than property reconstructors) to the setting of data privacy. Indeed, prior to this work it was not clear if property testers for Lipschitz property can be used at all in data privacy setting.

Let \mathcal{T} be the universe from which data sets are drawn where each data set has the same number of records. A function $f : \mathcal{T} \rightarrow \mathbb{R}$ is α -Lipschitz if for all pair of points $x, x' \in \mathcal{T}$ the following condition holds: $|f(x) - f(x')| \leq \alpha d_H(x, x')$ where d_H is the Hamming distance between x and x' (that is, $d_H(x, x')$ is the number of entries in which x and x' differ). To define Lipschitz tester, we define the notion of distance between functions f and g defined on the same (finite) domain \mathcal{T} under distribution \mathbf{Distr} as follows: $dist(f, g) \stackrel{\text{def}}{=} \Pr_{x \sim \mathbf{Distr}} [f(x) \neq g(x)]$. A Lipschitz tester gets an oracle access to function f , a distance parameter $\epsilon \in (0, 1]$. It accepts Lipschitz functions f and rejects with high probability functions f which are ϵ -far from Lipschitz property. Namely, functions f for which $\min dist(f, g) > \epsilon$, where the minimum is taken over all Lipschitz functions g . In this work, we extend the result of [JR11] to the setting of product distribution.

While \mathbf{Distr} is usually taken to be the uniform distribution in the property testing literature, in our setting it will be important to allow \mathbf{Distr} to be more general distribution. Taking \mathbf{Distr} to be something other than uniform distribution is challenging to investigate even for the special case of product distributions. Indeed, prior to this work the only positive result known for the product distribution setting is the work by [AC06] for monotonicity testing. For the setting where \mathbf{Distr} is an arbitrary unknown distribution there are exponential lower bounds on computational efficiency of the tester are known [HK07]. Above result is stated for functions with discrete range of the form $\delta\mathbb{Z}$.

In this paper, we show that one can use a Lipschitz property testing algorithm (*Liptest*) as a proxy for testing generalized differential privacy. The tester *Liptest* should be able to sample *efficiently* the data set according to a given probability distribution defined over domain of these data sets (see Definition 2.2). It has been shown that this additional requirement is sufficient to give strong privacy guarantees for the algorithm being tested. (For further details see Section 3.) Additionally, for practical applications, this tester should run efficiently, especially over the large data set domain.

With the above motivation in mind, we have designed such a Lipschitz tester with sub-linear time complexity (with respect to the domain size) for the hypercube domain $\mathcal{T} = \{0, 1\}^d$ with product distribution defined on data sets in \mathcal{T} . (For further details, we refer the reader to Section 4.) With this construction, we can test the privacy guarantees of an algorithm in time that is poly-logarithmic in domain size.

1.1 Related Work

In the last few years, various notions of data privacy have been proposed. Some of the most prominent are k -anonymity [Swe02], ℓ -diversity [MGKV06], differential privacy [DMNS06], noiseless privacy [BBG⁺11], natural differential privacy [BD12] and generalized differential privacy [BBG⁺11]. While ad-hoc notions like k -anonymity and ℓ -diversity being broken [GKS08], privacy community has pretty much converged to theoretical sound notions of privacy like differential privacy. In this paper, we work with the definition of generalized differential privacy (GDP), which is a generalization of differential privacy, noiseless privacy and natural differential privacy. The primary difference between GDP and the other related definitions is that it incorporates both the randomness in the underlying data set \mathcal{D} and the randomness of the Algorithm \mathcal{A} , where as other notions consider either the randomness of the data or the randomness of the algorithm.

In this paper, we design algorithms (\mathcal{A}_{test}) to test whether a given algorithm \mathcal{A} satisfies GDP. In all our algorithms, we assume that \mathcal{A} is given as a “white-box”, i.e., complete access to the source code of \mathcal{A} is provided. In this paper, all the instantiations of \mathcal{A}_{test} are probabilistic and use Lipschitz property testing algorithms as underlying tool set. On a related note, in the field of formal verifications there have been recent works [RP10] using which one can guarantee that a given algorithm \mathcal{A} satisfy differential privacy. The caveat of these kind of static analysis based algorithms is that it needs the source code for \mathcal{A} to be written in a type-safe language which is hard for a non-expert to adapt to.

One of the primary reason for considering the sublinear (with respect to the domain size) time Lipschitz testers is the large size of domain often encountered in the study of statistical privacy of databases. The property testers

([RS96b, GGR98b]) have been extensively studied for various approximation and decision problems. They are of particular interest because they usually have sublinear (in input size) running time which is of particular interests in the problem with large inputs. Some of the ideas and definitions in this paper have been taken from the work on distribution testing ([HK07, GS09, AC06]). Lipschitz property testers were introduced in [JR11] (which gave the explicit tester for the hypercube domain) and have since then been studied in [AJMR12b, AJMR12a] for the hypergrid domain. Recently [CS12] have proposed an optimal Lipschitz tester for the hypercube domain with the underlying distribution being uniform.

1.2 Our Contributions

- **Formulate testing of data privacy property as Lipschitz property testing:** In this paper we initiate the study of testing privacy properties of a given candidate algorithm \mathcal{A} . The specific privacy property that we test is *generalized differential privacy* (GDP) (see Definition 2.2). In order to design a tester for GDP property, we cast the problem of testing GDP property as a problem of testing Lipschitzness. (See Theorem 3.1.) The problem of testing Lipschitzness was initially proposed by [JR11].
- **Design a generic transformation to convert an Algorithm \mathcal{A} to its GDP variant:** We design a generic transformation to convert a candidate algorithm \mathcal{A} to its generalized differentially private variant. (See Theorem 3.5.)
- **New results for Lipschitz property testing:** In order to allow our privacy tester to be effective for a large class of data generating distributions, we extend the existing results of Lipschitz property testing to work with product distributions. We give the first efficient tester for the Lipschitz property for the hypercube domain which works for arbitrary product distribution. (See Theorem 4.1.) Previous works (even for other function properties) have mostly focused on the case of uniform distribution. To the best of our knowledge this is the only non-trivial positive result in property testing over arbitrary product distribution apart from the result of [AC06] on monotonicity testing.
- **Concrete instantiation of privacy testers based on old and new Lipschitz testers** We instantiate privacy tester using Lipschitz tester described in the previous item to get a concrete instantiation of privacy tester. This also leads to a concrete instantiation of Item 2 mentioned above. We also instantiate privacy testers based on known Lipschitz testers in the literature. This is summarized in Section 5.

1.3 Organization of the paper

In Section 2, we introduce the notions of privacy used in this paper, namely, differential privacy and generalized differential privacy. We also introduce the concepts of general property testing and the specific instantiation of Lipschitz property testing. In Section 3, we show the formal connection between testing of generalized differential privacy (GDP) and Lipschitz property testing. In Section 4, we state our new results of Lipschitz property testing over product distributions in the hypercube domain. In Section 5, we show that Lipschitz testers over the hypergrid domain can be used to test for GDP when the data sets are drawn uniformly from the hypergrid domain. Lastly, in Section 6 we conclude with discussions and open problems.

2 Preliminaries

2.1 Differential Privacy and Generalized Differential Privacy

In the last few years, differential privacy [DMNS06] has become a well-accepted notion of statistical data privacy in the data privacy community. At a high-level the definition of differential privacy implies that the output of a differentially private algorithm will be “almost” the same from an adversary’s perspective irrespective of an individual’s presence or absence in the underlying data set. The reason that it is a meaningful notion is because the

presence or absence of an individual in the data set does not affect the output of the algorithm “too much”. This high-level intuition can be formalized as below:

Definition 2.1 ((α, γ) -Differential Privacy [DMNS06, DKMN06]). *A randomized algorithm \mathcal{A} is (α, γ) -differentially private if for any two data sets \mathcal{D} and \mathcal{D}' drawn from a domain \mathcal{T} with $|\mathcal{D} \Delta \mathcal{D}'| = 1$ (Δ being the symmetric difference), and for all measurable sets $\mathcal{O} \subseteq \text{Range}(\mathcal{A})$ the following holds:*

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{O}] \leq e^\alpha \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{O}] + \gamma$$

In the above definition if $\gamma = 0$, we simply call it α -differential privacy. In this paper we intend to test if an algorithm \mathcal{A} is α -differentially private. In order to test the above, we mould the problem into a problem of testing Lipschitzness over the probability measure induced by Algorithm \mathcal{A} over a finite set S (see Section 3 for more discussion on this). Since, we want to test Lipschitzness efficiently with respect to the size of the set S , we will use a relaxed notion of differential privacy called *generalized differential privacy* (GDP) [BBG⁺11]. The main idea behind GDP is that it allows us to incorporate the randomness over the data generating distribution. This in turn allows us to incorporate the failure probability of the Lipschitzness testing algorithm (over the randomness of the data generating distribution). The definition of GDP below is a slight modification to the definition proposed in [BBG⁺11] and in most natural settings is stronger than [BBG⁺11].

Definition 2.2 ((α, γ, β) -Generalized Differential Privacy). *Let \mathbf{Dist} be the distribution over the space of all data sets drawn from domain \mathcal{T} . Let $W \subseteq \mathcal{T}$ be a set such that $\Pr_{\mathcal{D} \sim \mathbf{Dist}}[\mathcal{D} \in W] \leq \beta$. A randomized algorithm \mathcal{A} is (α, γ, β) -generalized differentially private (GDP) if for any pair data sets $\mathcal{D}, \mathcal{D}' \in \mathcal{T} \setminus W$ with $|\mathcal{D} \Delta \mathcal{D}'| = 1$ (Δ being the symmetric difference) and for all measurable sets $\mathcal{O} \subseteq \text{Range}(\mathcal{A})$ the following holds: $\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{O}] \leq e^\alpha \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{O}] + \gamma$, where the probability is over the randomness of the Algorithm \mathcal{A} .*

It is worth mentioning here that the above definition generalizes the *noiseless privacy* definition [BBG⁺11] and *natural differential privacy* definition [BD12] in the literature. While in both noiseless and natural differential privacy definitions the randomness is *solely* over the data generating distribution \mathbf{Dist} , in GDP the randomness is both over the data generating distribution and the randomness of the algorithm.

At a high-level what GDP says is that there exists a set W of “bad” data sets where (α, γ) -differential privacy condition does not hold. But the probability of drawing a data set \mathcal{D} (over the data generating distribution \mathbf{Dist}) from W is at most β (which is usually negligible in the problem parameters). In fact if we set $\beta = 0$, then we recover (α, γ) -differential privacy definition (see Definition 2.1) exactly. Similarly, it can be shown that under different choices of (α, γ, β) GDP implies both noiseless privacy and natural differential privacy.

2.2 Lipschitz Property Testing

In this work we show that efficiently testing whether an algorithm is (α, β, γ) -generalized differentially private reduces to the problem of testing (with high success probability over the probability measure induced by Algorithm \mathcal{A}) if the output is Lipschitz. (For further details see section, see section 3.)

Definition 2.3. *Given a function $f : \mathcal{T} \rightarrow \mathbb{R}$ from a metric space $(\mathcal{T}, d_{\mathcal{T}})$ to $(\mathbb{R}, d_{\mathbb{R}})$, where $d_{\mathcal{T}}$ and $d_{\mathbb{R}}$ denote the distance function on the domain \mathcal{T} and the range \mathbb{R} respectively. The function f is c -Lipschitz if $d_{\mathbb{R}}(f(x), f(y)) \leq c \cdot d_{\mathcal{T}}(x, y)$.*

Property testing ([GGR98b],[RS96b]) is a well studied area pertaining to randomized approximation algorithms for decision problems usually having sublinear time and query complexity. At one end of the spectrum, most of the work previously done in this area assume a uniform distribution over domain elements. The other end is to consider the setting where the distribution over the domain points is not known ([HK07]).

Here, we assume that the probability measure over domain elements is known and is not necessarily uniform. Although seemingly important, to the best of our knowledge, this is the first time that such a setting is explored in the Lipschitz property testing. To state our results, we will need the following notation.

Let \mathcal{P} (e.g. Lipschitzness in this case) be the property that needs to be tested over the range of function $f : D \rightarrow R$. We define the distance of the function f from \mathcal{P} as follows.

Definition 2.4. Let \mathcal{P} and \mathcal{T} be defined as above. The \mathcal{P} -distance between functions $f, g \in \mathcal{F}$ is defined by $\text{dist}_{\mathcal{P}}(f, g) \stackrel{\text{def}}{=} \Pr_{x \sim \mathcal{T}}\{f(x) \neq g(x)\}$. The \mathcal{P} -distance of a function f from property \mathcal{P} is defined as $\text{dist}_{\mathcal{P}}(f, \mathcal{P}) = \min_{g \in \mathcal{P}} \text{dist}_{\mathcal{P}}(f, g)$. We say that f is ϵ -far from a property \mathcal{P} if $\text{dist}_{\mathcal{P}}(f, \mathcal{P}) \geq \epsilon$.

We will need the notion of the image diameter of a function f for explaining our results, which, roughly speaking, is the difference between maximum and minimum values taken by f on domain \mathcal{T} .

Definition 2.5 (Image diameter). The image diameter of a function $f : \mathcal{T} \rightarrow \mathbb{R}$, denoted by $\text{ImD}(f)$, is the difference between the maximum and the minimum values attained by f , i.e., $\max_{x \in \mathcal{T}} f(x) - \min_{x \in \mathcal{T}} f(x)$.

3 Test for Generalized Differential Privacy

In this work we initiate the study of testing whether a given algorithm \mathcal{A} satisfies statistical data privacy guarantees. As a specific instantiation of the problem, we study the notion of generalized differential privacy (GDP) (see Definition 2.2). Roughly speaking, GDP guarantee ensures that the output of Algorithm \mathcal{A} when executed on data set \mathcal{D} does not depend “too much” on any one entry of \mathcal{D} . The term “too much” is formalized by three parameters α , γ and β , where the first two parameters (α and γ) depends on the randomness of the Algorithm \mathcal{A} and the parameter β depends on the randomness of the distribution **Distr** generating the data. We refer to the guarantee as (α, γ, β) -Generalized Differential Privacy (or simply (α, γ, β) -GDP).

Given an algorithm \mathcal{A} , we design a tester $\mathcal{A}_{\text{test}}$ with the following property: if the tester outputs **YES**, then Algorithm \mathcal{A} is (α, γ, β) -generalized differentially private where the parameters β and γ can be made arbitrarily small (at the cost of increased running time). If the tester outputs **NO**, then the Algorithm \mathcal{A} is *not* α -differentially private. We state this formally below.

Theorem 3.1 ($(\theta, \alpha, \gamma, \beta)$ -Privacy testing). Let **Liptest** be a θ -approximate Lipschitz tester (see Definition 3.2 below), let **Distr** be a distribution on the domain of datasets \mathcal{T} and let \mathcal{A} be an algorithm which on input $\mathcal{D} \sim \mathbf{Distr}$ outputs a value $\mathcal{A}(\mathcal{D})$ in the finite set Γ . Suppose there is an oracle $\mathcal{O}_{\mathcal{A}}$ which for every value $o \in \Gamma$ and for every $\mathcal{D} \in \mathcal{T}$ allows constant time access to the probability measure $\mu(\mathcal{A}(\mathcal{D}) = o)$ (where the measure is over the randomness of the algorithm \mathcal{A}). Then there exists a “testing” algorithm $\mathcal{A}_{\text{test}}$ which on input privacy parameters $\alpha, \beta \in (0, 1]$, failure probability parameter $\gamma \in (0, 1]$ and access to $\mathcal{O}_{\mathcal{A}}$ and **Distr** satisfies the following guarantee.

- (**soundness**) If Algorithm $\mathcal{A}_{\text{test}}$ outputs **NO**, then the candidate algorithm \mathcal{A} is *not* α -differentially private.
- (**completeness**) If Algorithm $\mathcal{A}_{\text{test}}$ outputs **YES**, then with probability at least $1 - \gamma$ the candidate algorithm \mathcal{A} is $(\alpha\theta, 0, \beta)$ -generalized differentially private.

The algorithm $\mathcal{A}_{\text{test}}$ uses **Liptest** as a subroutine and runs in time $O(|\Gamma| \cdot (\text{Run time of } \mathbf{Liptest}))$.

To prove Theorem 3.1, we show a new connection between testing $(\alpha, 0, \beta)$ -GDP and the problem of testing Lipschitz property. The study of testing Lipschitz property was initiated by [JR11]. We present an algorithm $\mathcal{A}_{\text{test}}$ for testing $(\alpha, 0, \beta)$ -GDP based on a generalization of Lipschitz tester presented in [JR11]. We formally define the (generalized) Lipschitz tester below where the definition differs from the standard property testing definition (example, as used in [JR11]) in two aspects: (i) we require Lipschitz testers to only distinguish between Lipschitz functions from functions which are far from θ -Lipschitz functions for some fixed $\theta \geq 1$ and (ii) we measure distance between functions (in particular, how “far” the function is from satisfying the property) with respect to a pre-defined probability measure **Distr** on the domain.

Definition 3.2 (θ -approximate Lipschitz tester). A θ -approximate Lipschitz tester **Liptest** is a randomized algorithm that gets as input: (i) oracle access to function $f : \mathcal{T} \rightarrow \mathbb{R}$; (ii) oracle access to independent samples from distribution **Distr** on \mathcal{T} and (iii) parameters $\epsilon, \gamma \in (0, 1]$. It outputs a **YES/NO** value and provides the following guarantee.

- If **Liptest** outputs **NO**, then with probability 1, the function f is **not** Lipschitz.
- If **Liptest** outputs **YES**, then with probability at least $1 - \gamma$, there exists a set $W \subseteq \mathcal{T}$ such that (i) the input function f is θ -Lipschitz on the domain $\mathcal{T} \setminus W$ and (ii) $\Pr_{\mathcal{D} \sim \mathbf{Distr}}[\mathcal{D} \in W] \leq \epsilon$.

We remark that setting $\theta = 1$ and **Distr** to be the uniform distribution on \mathcal{T} recovers the standard definition of property tester (in our case, Lipschitz tester as defined in [JR11]).

In Section 3.2, we show that one can extend the connection between GDP and Lipschitz testing to design an algorithm $\mathcal{A}_{\text{privGen}}$ which converts the candidate algorithm \mathcal{A} in to a (α, γ, β) -generalized differentially private algorithm.

3.1 (Generalized) Differential Privacy as Lipschitz Property over a Probability Measure

Consider the domain of the data sets \mathcal{T} to be a finite set and assume that (the randomized) Algorithm \mathcal{A} , whose privacy property is to be tested, maps a data set $\mathcal{D} \in \mathcal{T}$ to another finite set Γ , i.e. any output of \mathcal{A} is always an element in Γ . Now let us look at the privacy guarantee of GDP (see Definition 2.2). Ignoring the parameters β and γ , the privacy guarantee suggests that for any pair of neighboring data sets $\mathcal{D}, \mathcal{D}' \in \mathcal{T}$ (drawn from the distribution **Distr**) and any $o \in \Gamma$, the following is true:

$$e^{-\alpha} \mu(\mathcal{A}(\mathcal{D}') = o) \leq \mu(\mathcal{A}(\mathcal{D}) = o) \leq e^{\alpha} \mu(\mathcal{A}(\mathcal{D}') = o) \quad (1)$$

The measure μ is the probability induced by the randomness of the Algorithm \mathcal{A} . Taking logarithm of (1), we get

$$|\log \mu(\mathcal{A}(\mathcal{D}) = o) - \log \mu(\mathcal{A}(\mathcal{D}') = o)| \leq \alpha \quad (2)$$

We will use the following formulation of (2): $|\frac{1}{\alpha} \log \mu(\mathcal{A}(\mathcal{D}) = o) - \frac{1}{\alpha} \log \mu(\mathcal{A}(\mathcal{D}') = o)| \leq d_H(\mathcal{D}, \mathcal{D}')$, where d_H is the Hamming metric. Now, if we view the expression $\frac{1}{\alpha} \log \mu(\mathcal{A}(\mathcal{D}) = o)$ as a function $\lambda_o : \mathcal{T} \rightarrow \mathbb{R}$ defined by setting $\lambda_o(\mathcal{D}) = \frac{1}{\alpha} \log \mu(\mathcal{A}(\mathcal{D}) = o)$, then we get the following condition: $|\lambda_o(\mathcal{D}) - \lambda_o(\mathcal{D}')| \leq d_H(\mathcal{D}, \mathcal{D}')$. This condition is exactly the Lipschitzness guarantee for λ_o under the Hamming metric. Using this observation we state the following meta-algorithm $\mathcal{A}_{\text{test}}$ (Algorithm 1) to test whether given Algorithm \mathcal{A} is $(\alpha, 0, \beta)$ -generalized differentially private. In Algorithm 1 (Algorithm $\mathcal{A}_{\text{test}}$), we use a black box Lipschitz property tester **Liptest**. Later in the paper we instantiate **Liptest** with a specific testing algorithms.

Algorithm 1 $\mathcal{A}_{\text{test}}$: Generalized Differential Privacy (GDP) tester

Require: Algorithm \mathcal{A} , data generating distribution **Distr**, data domain \mathcal{T} , output range Γ , privacy parameters (α, β) and failure parameter γ

- 1: $flag \leftarrow \text{FALSE}$
 - 2: Let **Liptest** be a θ -approximate Lipschitz tester defined in Definition 3.2.
 - 3: **for all** values $o \in \Gamma$ **do**
 - 4: Define function $\lambda_o : \mathcal{T} \rightarrow \mathbb{R}$ by setting $\lambda_o(\mathcal{D}) = \frac{1}{\alpha} \log \mu(\mathcal{A}(\mathcal{D}) = o)$.
 - 5: Run **Liptest** on λ_o with *proximity* parameter $\frac{\beta}{\Gamma}$ and failure probability parameter $\frac{\gamma}{|\Gamma|}$.
 - 6: If **Liptest** outputs **NO**, then $flag \leftarrow \text{TRUE}$
 - 7: **end for**
 - 8: If $flag = \text{FALSE}$, then output **YES**, otherwise output **NO**
-

At a high-level Algorithm \mathcal{A}_{test} does the following. For each possible output $o \in \Gamma$, it defines a function table λ_o (with the domain \mathcal{T}). It then invokes the Lipschitz testing algorithm **Liptest** to test λ_o for Lipschitzness property. If for every output $o \in \Gamma$, **Liptest** outputs **YES**, then \mathcal{A}_{test} outputs affirmative, and outputs negative otherwise.

3.1.1 Proof of Theorem 3.1

The claim about the running time of Algorithm \mathcal{A}_{test} stated in Theorem 3.1 follows directly from the definition of Algorithm \mathcal{A}_{test} (Algorithm 1). We state and prove the soundness and completeness guarantees of Theorem 3.1 separately as Claim 3.3 and Claim 3.4 respectively below.

Claim 3.3 (Soundness guarantee). *If Algorithm \mathcal{A}_{test} (Algorithm 1) outputs **NO**, then the candidate algorithm \mathcal{A} is not α -differentially private.*

Proof. If Algorithm \mathcal{A}_{test} outputs a **NO**, then there exists an $o \in \Gamma$ such that **Liptest** outputs **NO** on λ_o . By definition of **Liptest** (see Definition 3.2), we get that λ_o is not Lipschitz. In other words, we have, $|\lambda_o(\mathcal{D}) - \lambda_o(\mathcal{D}')| = |\frac{1}{\alpha} \log \mu(\mathcal{A}(\mathcal{D}) = o) - \frac{1}{\alpha} \log \mu(\mathcal{A}(\mathcal{D}') = o)| > 1$. Therefore, either $\mu(\mathcal{A}(\mathcal{D}) = o) > e^\alpha \mu(\mathcal{A}(\mathcal{D}') = o)$ or $\mu(\mathcal{A}(\mathcal{D}) = o) < e^{-\alpha} \mu(\mathcal{A}(\mathcal{D}') = o)$, as required. \square

Claim 3.4 (Completeness guarantee). *If Algorithm \mathcal{A}_{test} (Algorithm 1) outputs **YES**, then with probability at least $1 - \gamma$ (over the randomness of **Liptest**), the candidate algorithm \mathcal{A} is $(\alpha\theta, 0, \beta)$ -generalized differentially private.*

Proof. If Algorithm \mathcal{A} outputs **YES**, then by the union bound it follows that with probability at least $1 - \gamma$, the following condition holds for every $o \in \Gamma$: There exists a set $W_o \subseteq \mathcal{T}$ such that (i) λ_o satisfies θ -Lipschitz condition for every $\mathcal{D}, \mathcal{D}' \in \mathcal{T} \setminus W_o$ and (ii) $\Pr_{x \sim \text{Distr}}[x \in W_o] < \frac{\beta}{|\Gamma|}$.

Let $W = \bigcup_{o \in \Gamma} W_o$. We show that with probability at least $1 - \gamma$ (over the randomness of **Liptest**), the following holds: algorithm \mathcal{A} satisfies $\alpha\theta$ -differential privacy condition on the set $\mathcal{T} \setminus W$ and $\Pr_{\mathcal{D} \sim \text{Distr}}[\mathcal{D} \in W] \leq \beta$.

Condition (i) above implies that for every $o \in \Gamma$, λ_o is θ -Lipschitz on $\mathcal{T} \setminus W$. Therefore, we get the following for every neighboring pairs of data sets $\mathcal{D}, \mathcal{D}' \in \mathcal{T} \setminus W$.

$$\begin{aligned} |\lambda_o(\mathcal{D}) - \lambda_o(\mathcal{D}')| &\leq \theta \\ \Rightarrow \left| \frac{1}{\alpha} \log \mu(\mathcal{A}(\mathcal{D}) = o) - \frac{1}{\alpha} \log \mu(\mathcal{A}(\mathcal{D}') = o) \right| &\leq \theta \\ \Rightarrow e^{-\alpha\theta} &\leq \frac{\mu(\mathcal{A}(\mathcal{D}) = o)}{\mu(\mathcal{A}(\mathcal{D}') = o)} \leq e^{\alpha\theta} \end{aligned}$$

Also, using Condition (ii) and the union bound over all $o \in \Gamma$, we get the following.

$$\Pr_{\mathcal{D} \sim \text{Distr}}[\mathcal{D} \in W] \leq \sum_{o \in \Gamma} \Pr_{\mathcal{D} \sim \text{Distr}}[\mathcal{D} \in W_o] \leq \beta.$$

Since Conditions (i) and (ii) both hold with probability at least $1 - \gamma$ (over the randomness of **Liptest**), we get the desired claim. \square

3.2 Application of GDP tester to ensure privacy for the output of a given candidate algorithm

In this section we will demonstrate how one can use Algorithm \mathcal{A}_{test} (Algorithm 1) designed in the previous section to guarantee (α, β, γ) -generalized differential privacy to the output produced by a candidate Algorithm \mathcal{A} . The details are given in Algorithm 2. The theoretical guarantees for Algorithm 2 are given below.

Theorem 3.5 ($(\theta, \alpha, \gamma, \beta)$ -generalized differentially private mechanism). *Let $\mathcal{Liptest}$ be a θ -approximate Lipschitz tester (see Definition 3.2) used in the testing algorithm \mathcal{A}_{test} (Algorithm 1). Under the same assumptions of Theorem 3.1, following are true for Algorithm $\mathcal{A}_{privGen}$ (Algorithm 2).*

- **(privacy)** Algorithm $\mathcal{A}_{privGen}$ (Algorithm 2) is $(\alpha\theta, \beta, \gamma)$ -generalized differentially private (GDP).
- **(utility)** If the candidate Algorithm \mathcal{A} is α -differentially private, then Algorithm $\mathcal{A}_{privGen}$ (Algorithm 2) always produces the output $\mathcal{A}(\mathcal{D})$.

Algorithm 2 $\mathcal{A}_{privGen}$: Generalized differentially private mechanism

Require: Data set \mathcal{D} , candidate algorithm \mathcal{A} , testing algorithm \mathcal{A}_{test} , data generating distribution \mathbf{Distr} , data domain \mathcal{T} , output set Γ , privacy parameters (α, β, γ)

- 1: Run \mathcal{A}_{test} with parameters $\mathcal{A}, \mathbf{Distr}, \mathcal{T}, \Gamma$, privacy parameters (α, β) , and failure parameter γ
 - 2: If \mathcal{A}_{test} outputs **YES**, then output $\mathcal{A}(\mathcal{D})$, output **FAILURE** otherwise
-

3.2.1 Proof of Theorem 3.5

The proof of Theorem 3.5 follows from the two claims below.

Claim 3.6 (Privacy). *Algorithm $\mathcal{A}_{privGen}$ (Algorithm 2) is $(\alpha\theta, \gamma, \beta)$ -generalized differentially private (GDP).*

Proof. First note that from Claim 3.4, it follows that if Algorithm \mathcal{A}_{test} (Algorithm 1) outputs **YES**, then w.p. $\geq 1 - \gamma$, the candidate algorithm \mathcal{A} is $(\alpha\theta, 0, \beta)$ -GDP. Now to complete the proof, we provide the following argument.

- **Case 1 [Algorithm 2 outputs $\mathcal{A}(\mathcal{D})$]:** We define event Ev to be the following: For every $o \in \Gamma$ there exists a set $W_o \subseteq \mathcal{T}$ such that (i) λ_o satisfies θ -Lipschitz condition for every $\mathcal{D}, \mathcal{D}' \in \mathcal{T} \setminus W_o$ and (ii) $\Pr_{x \sim \mathbf{Distr}}[x \in W_o] < \beta$. As implied by the GDP guarantee, event Ev holds with probability $1 - \gamma$. Hence, we have the following for all $o \in \Gamma \cup \{\mathbf{FAILURE}\}$

$$\begin{aligned}
\Pr[\mathcal{A}_{privGen}(\mathcal{D}) = o] &\leq \Pr[\mathcal{A}_{privGen}(\mathcal{D}) = o | Ev] \Pr[Ev] + \Pr[\bar{Ev}] \\
&\leq e^{\alpha\theta} \Pr[\mathcal{A}_{privGen}(\mathcal{D}') = o | Ev] \Pr[Ev] + \gamma \\
&\leq e^{\alpha\theta} \Pr[\mathcal{A}_{privGen}(\mathcal{D}') = o \wedge Ev] + \gamma \\
&\leq e^{\alpha\theta} \Pr[\mathcal{A}_{privGen}(\mathcal{D}') = o] + \gamma
\end{aligned}$$

- **Case 2[Algorithm 2 outputs **FAILURE**]:** In this case, the output is trivially (α, γ, β) -generalized differentially private since the output (i.e., **FAILURE**) is independent of the data set \mathcal{D} .

With this the proof is complete. □

Claim 3.7 (Utility). *If the candidate Algorithm \mathcal{A} is α -differentially private, then Algorithm $\mathcal{A}_{privGen}$ (Algorithm 2) always produces the output $\mathcal{A}(\mathcal{D})$.*

The proof of the above claim follows from the fact that if the candidate algorithm \mathcal{A} is α -differentially private, then \mathcal{A}_{test} will always output **YES**.

4 Lipschitz Property Testing over Hypercube domain

In this section, we present a $(1 + \delta)$ -approximate Lipschitz tester (see Definition 3.2) for functions defined on $\mathcal{T} = \{0, 1\}^d$ where the notion of distance is with respect to any product distribution. Specifically, the points in the data set are distributed according to the product distribution $\Pi = \text{Ber}(p_1) \times \text{Ber}(p_2) \times \dots \times \text{Ber}(p_d)$ where $\text{Ber}(p)$ denotes the Bernoulli distribution with probability p . For any vertex $x = (x_1, x_2, \dots, x_d) \in \mathcal{T}$, $x_i = 1$ with probability p_i and 0 with probability $1 - p_i$. Each vertex in $x \in \mathcal{T}$ has an associated probability mass $p_x = p_{i_1} \cdot p_{i_2} \cdots p_{i_k} \cdot (1 - p_{j_1}) \cdot (1 - p_{j_2}) \cdots (1 - p_{j_{d-k}})$ where k is the hamming weight of x , also denoted by $H(x)$ and i_1, i_2, \dots, i_k denote the indices of x with bit-value 1.

In this section, we prove the following theorem which gives a 1-approximate Lipschitz tester for $\delta\mathbb{Z}$ -valued functions. A function is $\delta\mathbb{Z}$ valued if it produces outputs in integral multiples of δ .

Theorem 4.1. *Let $\mathcal{T} = \{0, 1\}^d$ be the domain from which the data set are drawn according to a product probability distribution $\Pi = \text{Ber}(p_1) \times \text{Ber}(p_2) \times \dots \times \text{Ber}(p_d)$. The Lipschitz property of functions $f : \mathcal{T} \rightarrow \delta\mathbb{Z}$ on these data sets can be tested non-adaptively and with one sided error probability ω in $O(\frac{d \cdot \min\{d, \text{ImD}(f)\}}{\delta(\epsilon - d^2\delta)} \ln(\frac{2}{\omega}))$ time for $\delta \in (0, 1]$. Here ImD is the image diameter defined in Definition 2.5.*

Following is an easy corollary of the above giving a $(1 + \delta)$ -approximate Lipschitz tester for \mathbb{R} -valued functions.

Corollary 4.2 (of Theorem 4.1). *Let $\mathcal{T} = \{0, 1\}^d$ be the domain from which the data set are drawn according to a product probability distribution $\Pi = \text{Ber}(p_1) \times \text{Ber}(p_2) \times \dots \times \text{Ber}(p_d)$. There is an algorithm that on input parameters $\delta \in (0, 1], \epsilon \in (0, 1), d$ and oracle access to a function $f : \{0, 1\}^d \rightarrow \mathbb{R}$ has the following behavior: It accepts if f is Lipschitz and rejects with probability at least $1 - \omega$ if f is ϵ -far (with respect to the distribution Π) from $(1 + \delta)$ -Lipschitz and runs in $O(\frac{d \cdot \min\{d, \text{ImD}(f)\}}{\delta(\epsilon - d^2\delta)} \ln(\frac{2}{\omega}))$ time. Here ImD is the image diameter defined in Definition 2.5.*

The proof of above theorem and corollary appears in Section 4.1. To state the proof we need the following technical result.

We define a distribution on edges of the hypercube where the probability mass of an edge $\{x, y\}$ is given by $\frac{p_x + p_y}{d}$. Note that $\sum_{(x,y) \in E(H_d)} \frac{(p_x + p_y)}{d} = 1$. Thus the probability distribution (we call it D_E henceforth) on the edges defined above is consistent. Our tester is based on detecting violated edges (that is, edges which violate Lipschitz property) sampled from distribution D_E . Our main technical lemma (Lemma 4.3) gives a lower bound on the probability of sampling a violated edge according to distribution D_E for a function that is ϵ -far from Lipschitz. (Recall that ϵ -far is measured with respect to the distribution Π .)

Lemma 4.3. *Let function $f : \{0, 1\}^d \rightarrow \delta\mathbb{Z}$ be ϵ -far from Lipschitz. Then*

$$\sum_{(x,y) \in V(f)} \frac{(p_x + p_y)}{d} \geq \frac{\delta(\epsilon - d^2\delta)}{d \cdot \text{ImD}(f)}$$

Here ImD is the image diameter defined in Definition 2.5.

We prove the above lemma in section 4.2.1.

4.1 Lipschitz tester

In this section we prove Theorem 4.1 and Corollary 4.2. We first present the algorithm stated in Theorem 4.1.

Proof of Theorem 4.1. First observe that if input function f is Lipschitz then the Algorithm 3 always accepts. This is because a Lipschitz function f has image diameter (see Definition 2.5) at most d (and hence cannot be rejected in Step 4. Moreover, it does not have any violated edges (and hence cannot be rejected in Step 6). Next consider the case when f is ϵ -far from Lipschitz. Towards this we first extend Claim 3.1 of [JR11] about sample diameter r to our setting where the distance (in particular, the notion of ϵ -far) is measured with respect to product distribution.

Algorithm 3 Lipschitz Tester

Require: Data domain $\mathcal{T} = \{0, 1\}^d$, product distribution on data set $\Pi = \text{Ber}(p_1) \times \text{Ber}(p_2) \times \dots \times \text{Ber}(p_d)$, failure probability parameter ω , \mathcal{P} -distance parameter ϵ' , discretization parameter δ

- 1: Set $\epsilon = \epsilon' - d^2\delta$.
 - 2: Sample $\lceil \frac{2}{\epsilon} \ln(\frac{2}{\omega}) \rceil$ vertices z_1, z_2, \dots, z_t independently from \mathcal{T} according to the distribution Π
 - 3: Let $r = \max_{i=1}^t f(z_i) - \min_{i=1}^t f(z_i)$
 - 4: If $r > d$, reject
 - 5: Sample $\lceil \frac{dr}{\delta\epsilon} \ln(\frac{2}{\omega}) \rceil$ edges independently with each edge (x, y) picked with probability $\frac{(p_x + p_y)}{d}$ from the hypercube \mathcal{T}
 - 6: If any of the sampled edges are violated, then reject, else accept
-

Claim 4.4. *The steps 1. and 2. of the tester outputs $r \in \delta\mathbb{Z}$ such that $r \leq \text{ImD}(f)$ and with probability at least $1 - \frac{\omega}{2}$ (failure probability at most $\frac{\omega}{2}$), f is ϵ -close to having diameter that is at most r .*

Proof. Sort the points in $\{0, 1\}^d$ according the function value in non-decreasing order. Let L be the first ℓ -points such that their *probability mass* sums up to $\frac{\epsilon}{2}$ and R be the set of last ℓ' points such that their *probability mass* sums up to $\frac{\epsilon}{2}$. The rest of the proof is very similar to the proof of Claim 3.1 in [JR11], so we omit the details here. \square

Having established Claim 4.4, rest of the proof is identical to [JR11] and we omit the details. \square

Proof of Corollary 4.2. It is identical to the proof of Corollary 1.2 in [JR11] and we omit the details. \square

4.2 Repair Operator and Proof of Lemma 4.3

We show a transformation of an arbitrary function $f : \{0, 1\}^d \rightarrow \delta\mathbb{Z}$ into Lipschitz function by changing f on certain points, whose probability mass is related to the probability mass (with respect to D_E) of the violated edges of \mathcal{T} . This is achieved by repairing one dimension of \mathcal{T} at a time as explained henceforth. To achieve this, we define an **asymmetric** version of the basic operator of [JR11]. The operator redefines function values so that it reduces the gap asymmetrically according to the Hamming weights (and probability masses in-turn) of the endpoints of the violated edge. This is the main difference from previous approaches ([JR11], [AJMR12b]) which do not work if applied directly, because of the varying probability masses of the vertices with respect to the Hamming weight. We first define the building block of the repair operator which is called the asymmetric basic operator.

Definition 4.5 (Asymmetric basic operator). *Given $f : \{0, 1\}^d \rightarrow \delta\mathbb{Z}$, for each violated edge $\{x, y\}$ along dimension i , where $f(x) < f(y) - 1$, define B_i as follows.*

1. *If $H(x) > H(y)$, then $B_i[f](x) = f(x) + (1 - p_i)\delta$ and $B_i[f](y) = f(y) - p_i\delta$*
2. *If $H(x) < H(y)$, then $B_i[f](x) = f(x) + p_i\delta$ and $B_i[f](y) = f(y) - (1 - p_i)\delta$*

Now we define the repair operator.

Definition 4.6 (Repair operator). *Given $f : \{0, 1\}^d \rightarrow \delta\mathbb{Z}$, $A_i[f](x)$ is obtained from f by several applications of the asymmetric basic operator (see Definition 4.5) B_i along dimension i followed by a single application of the rounding operator. Specifically, let f' be the function obtained from f by applying B_i repeatedly until there are no violated edges along the i -th dimension. Then, $A_i[f]$ is defined to be $\mathbf{R}[f']$ where the rounding operator \mathbf{R} rounds the function values to the closest $\delta\mathbb{Z}$ -valued function.*

In effect, we have the following picture for the repair operation.

$$f = f_0 \xrightarrow{\mathbf{R} \circ B_1^{\lambda_1}} f_1 \xrightarrow{\mathbf{R} \circ B_2^{\lambda_2}} f_2 \rightarrow \dots \rightarrow f_{d-1} \xrightarrow{\mathbf{R} \circ B_d^{\lambda_d}} f_d.$$

Now we define a measure called violation score which will be used to show the progress of repair operation. As shown later, the violation score is approximately preserved along any dimension $j \neq i$ when we apply the repair operator to repair the edges along dimension i . Note that the violation score closely resembles the violation score in [JR11] except that it depends on the function value as well as the probability masses of the end-points of the edge.

Definition 4.7. *The violation score of an edge with respect to function f , denoted by $vs(\{x, y\})$, is $\max(0, (p_x + p_y)(|f(x) - f(y)| - 1))$. The violation score along dimension i , denoted by $VS^i(f)$, is the sum of violation scores of all edges along dimension i*

The violation score of an edge $\{x, y\}$ is positive iff it is violated and violation score of a $\delta\mathbb{Z}$ valued function is contained in the interval $[\delta(p_x + p_y), \text{ImD}(f)(p_x + p_y)]$. Let $V^i(f)$ denote be the set of edges along dimension i violated by f . Then

$$\delta \cdot \sum_{\{x, y\} \in V^i(f)} (p_x + p_y) \leq VS^i(f) \leq \sum_{\{x, y\} \in V^i(f)} (p_x + p_y) \cdot \text{ImD}(f) \quad (3)$$

Lemma 4.9 shows that A_i does not increase the violation score in dimensions other than i more than the additive value of δ . The lemma makes use of the following claim.

Claim 4.8 (Rounding is safe). *Given $a, b \in \mathbb{R}$ satisfying $|a - b| \leq 1$, let a' (respectively, b') be the value obtained by rounding a (respectively, b) to the closest $\delta\mathbb{Z}$ integer. Then $|a' - b'| \leq 1$.*

Proof. Assume without loss of generality $a \leq b$. For $x \in \mathbb{R}$, let $\lfloor x \rfloor_\delta$ be the largest value in $\delta\mathbb{Z}$ not greater than x . Observe that $a' \in \{\lfloor a \rfloor_\delta, \lfloor a \rfloor_\delta + \delta\}$. Using the fact that $\lfloor a \rfloor_\delta \leq b' \leq \lfloor a \rfloor_\delta + 1 + \delta$, we see that if $a' = \lfloor a \rfloor_\delta + \delta$ then $|b' - a'| \leq 1$ always holds. Therefore, assume $a' = \lfloor a \rfloor_\delta$. This can happen only if $a \leq \lfloor a \rfloor_\delta + \delta/2$. The latter implies $b \leq \lfloor a \rfloor_\delta + 1 + \delta/2$ (using the fact that $b - a \leq 1$). That is $b' \neq \lfloor a \rfloor_\delta + 1 + \delta$. In other words, $b' \leq \lfloor a \rfloor_\delta + 1$ again implying $b' - a' \leq 1$, as required. \square

Lemma 4.9. *For all $i, j \in [d]$, where $i \neq j$, and every function $f : \{0, 1\}^d \rightarrow \delta\mathbb{Z}$, the following holds.*

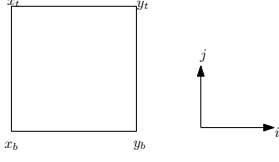
- **(progress)** *Applying the repair operator A_i does not introduce new violated edges in dimension j if the dimension j is violation free, i.e. $VS_j(f) = 0 \Rightarrow VS_j(A^i[f]) = 0$.*
- **(accounting)** *Applying the repair operator A_i does not increase the violation score in dimension j by more than δ , i.e. $VS_j(A^i[f]) \leq VS_j(f) + \delta$.*

Proof. Let f' be the function obtained from f by applying B_i repeatedly until there are no violated edges along the i -th dimension. We prove the following stronger claim to prove the lemma.

Claim 4.10. $VS_j(f') \leq VS_j(f)$.

We prove the above claim momentarily but first prove the lemma using the above claim. The function $A_i[f]$ is obtained by rounding the values of f' to the closest $\delta\mathbb{Z}$ values. Since rounding can never create new edge violations by Claim 4.8, we immediately get the first part of the lemma. The second part follows from the observation that the rounding step modifies each function value by at most $\delta/2$. Correspondingly, the violation score of an edge along the j -th dimension changes by at most $2 \cdot (\delta/2) \cdot (p_u + p_v)$ where the factor 2 comes because both endpoints of an edge may be rounded. Summing over all edges in the j -th dimension, we get, increase in violation score $\leq \sum_{\{u, v\}} \delta \cdot (p_u + p_v) = \delta$ where the last equality holds because edges along the j -th dimension form a perfect matching and therefore the probabilities $p_u + p_v$ sum to 1.

Proof of Claim 4.10. Following the proof outline of a similar proof in [JR11], we show that application of the asymmetric basic operator in dimension i does not increase the violation score in dimension $j \neq i$. Standard arguments [GGL⁺00, DGL⁺99, JR11, AJMR12b] show that it is enough to analyze the effect of applying B_i



on one fixed disjoint square formed by adjacent edges that cross dimensions i and j . (This is because edges along dimensions i and j form disjoint squares in the hypercube. So having established Claim 4.10 for one fixed square of the hypercube, the full claim follows by summing up the inequalities over all such squares.) Consider the two dimensional function $f : \{x_b, x_t, y_b, y_t\} \rightarrow \delta\mathbb{Z}$ where $\{x_b, x_t, y_b, y_t\}$ are positioned such that $H(y_t) = H(x_t) + 1 = H(y_b) + 1 = H(x_b) + 2$ where $H(x_b)$ denotes the hamming weight of x_b . Assume that the basic operator is applied along the dimension i . We show that the violation score along dimension j does not increase. Assume that the violation score along edge $\{x_b, x_t\}$ increases. First, assume that the $B_i[f](x_t) > B_i[f](x_b)$. (The other case is very similar and we will prove it later.) Then B_i increases $f(x_t)$ and/or decreases $f(x_b)$. Assume that B_i increases $f(x_t)$. (The other case is symmetrical.) This implies that $\{x_t, y_t\}$ is violated and $f(x_t) < f(y_t)$. Let $f_k(x)$ (resp. $f_k(y)$) denote the value of $f(x)$ (resp. $f(y)$) after k applications of B_i on an edge (x, y) , for an integer $k \geq 0$. If (x, y) is violated after $k - 1$ applications of the basic operator, then $f_k(x) = f_{k-1}(x) + p_i\delta$ and $f_k(y) = f_{k-1}(y) - (1 - p_i)\delta$ else $f_k(x) = f_{k-1}(x)$ and $f_k(y) = f_{k-1}(y)$. We will study the effect of applying B_i on (x_t, y_t) multiple (say $\lambda \geq 1$) times. Recall that the repair operator is applied only if the edge is violated. This means that

$$\begin{aligned} f_{\lambda-1}(x_t) &< f_{\lambda-1}(y_t) - 1 \\ \Rightarrow f(x_t) + (\lambda - 1)p_i\delta &< f(y_t) - (\lambda - 1)(1 - p_i)\delta - 1 \\ \Rightarrow f(x_t) + (\lambda - 1)\delta + 1 &< f(y_t) \\ \Rightarrow f(x_t) + \lambda\delta + 1 &\leq f(y_t) \end{aligned}$$

The second inequality follows from the observation that since the edge is being corrected in the λ^{th} application, it must have been corrected in all previous applications as well. The last inequality follows from the fact that f is a $\delta\mathbb{Z}$ -valued function and $\frac{1}{\delta}$ is an integer. We subtract $(1 - p_i)(\lambda - 1)\delta$ from both sides in the above inequality and do some rearrangement to achieve the following.

$$\begin{aligned} f(y_t) - (1 - p_i)(\lambda - 1)\delta &\geq f(x_t) + \lambda\delta + 1 - (1 - p_i)(\lambda - 1)\delta \\ \Rightarrow f(y_t) - (1 - p_i)(\lambda - 1)\delta &\geq f(x_t) + (\lambda - 1)p_i\delta + 1 + \delta \\ \Rightarrow f_{\lambda-1}(y_t) &\geq f_{\lambda-1}(x_t) + 1 + \delta \end{aligned}$$

The above inequality is crucial for the remaining proof of the lemma 4.3. Now consider the cases when either the bottom edge is also violated or is not violated.

If the bottom edge is not violated then we have $f_{\lambda-1}(x_b) \geq f_{\lambda-1}(y_b) - 1$ and $f_{\lambda-1}(x_b)$ and $f_{\lambda-1}(y_b)$ are not modified by the basic operator. Since $vs(\{x_t, x_b\})$ increases, $f_{\lambda-1}(x_t) > f_{\lambda-1}(x_b) + 1 - p_i\delta$. Combining the above inequalities, we get $f_{\lambda-1}(y_t) \geq f_{\lambda-1}(x_t) + 1 + \delta > f_{\lambda-1}(x_b) + 2 + (1 - p_i)\delta \geq f_{\lambda-1}(y_b) + 1 + (1 - p_i)\delta > f_{\lambda-1}(y_b) + 1$. Thus the violation score increases along $\{x_t, x_b\}$ by $(p_{x_b} + p_{x_t})p_i\delta$ and decreases along $\{y_b, y_t\}$ by $(p_{y_b} + p_{y_t})(1 - p_i)\delta = (p_{x_b} + p_{x_t})\left(\frac{p_i}{1 - p_i}\right)(1 - p_i)\delta$ which is same as $(p_{x_b} + p_{x_t})p_i\delta$, keeping the violation score along the dimension j unchanged.

If the bottom edge is violated, then the increase in $vs(\{x_b, x_t\})$ implies that $f_{\lambda-1}(x_b)$ must decrease (after application of B_i) by $p_i\delta$ (since $H(x_b) < H(y_b)$) implying $f_{\lambda-1}(y_b) + 1 < f_{\lambda-1}(x_b)$. Therefore $f_{\lambda-1}(x_t) + p_i\delta > f_{\lambda-1}(x_b) + 1 - p_i\delta$ or $f_{\lambda-1}(x_t) > f_{\lambda-1}(y_t) + 1 - 2p_i\delta$. Therefore $f_{\lambda-1}(y_t) > f_{\lambda-1}(x_t) + 1 > f(x_b) + 2 - 2p_i\delta \geq f(y_b) + 3 - 2p_i\delta + \delta \geq f(y_b) + 1 + \delta$. The last inequality is true since $\delta \leq 1$ and $p_i \leq 1$. Thus, $vs(\{x_t, x_b\})$

increases by at most $(p_{x_b} + p_{x_t})2p_i\delta$ while $vs(\{y_t, y_b\})$ decreases by $(p_{y_t} + p_{y_b})2(1 - p_i)\delta = (p_{x_b} + p_{x_t})2p_i\delta$, ensuring that violation score along the vertical dimension does not increase.

Now we turn to the case when $B_i[f](x_t) < B_i[f](x_b)$. By the arguments very similar to the first case, it can be proved that $f_{\lambda-1}(x_t) \geq f_{\lambda-1}(y_t) + 1 + \delta$ and the application of basic operator decreases $f(x_t)$ by $p_i\delta$ and increases $f(y_t)$ by $(1 - p_i)\delta$.

If the bottom edge is not violated then $f_{\lambda-1}(y_b) \geq f_{\lambda-1}(x_b) - 1$ and $f_{\lambda-1}(x_b)$ and $f_{\lambda-1}(y_b)$ are not modified by the basic operator. Since $vs(\{x_t, x_b\})$ increases, $f_{\lambda-1}(x_b) > f_{\lambda-1}(x_t) + 1 - p_i\delta$. Combining the above inequalities, we get $f_{\lambda-1}(y_b) \geq f_{\lambda-1}(x_b) - 1 > f(x_t) - p_i\delta \geq f(y_t) + 1 + \delta(1 - p_i)$. Thus the violation score increases along $\{x_t, x_b\}$ by $(p_{x_b} + p_{x_t})p_i\delta$ and decreases along $\{y_b, y_t\}$ by $(p_{y_b} + p_{y_t})(1 - p_i)\delta = (p_{x_b} + p_{x_t})\left(\frac{p_i}{1 - p_i}\right)(1 - p_i)\delta$ which is same as $(p_{x_b} + p_{x_t})p_i\delta$, keeping the violation score along the dimension j unchanged.

If the bottom edge is violated, then the increase in $vs(\{x_b, x_t\})$ implies that $f_{\lambda-1}(x_b)$ must increase implying $f_{\lambda-1}(y_b) > f_{\lambda-1}(x_b) + 1$. Therefore, the increase in $vs\{x_b, x_t\}$ implies that $f_{\lambda-1}(x_b) + p_i\delta > f_{\lambda-1}(x_t) - p_i\delta + 1$ or $f_{\lambda-1}(x_b) > f_{\lambda-1}(x_t) - 2p_i\delta + 1$. Combining the above inequalities, we get $f_{\lambda-1}(y_b) > f_{\lambda-1}(x_b) + 1 > f_{\lambda-1}(x_t) - 2p_i\delta + 2 \geq f_{\lambda-1}(y_t) + 3 + \delta - 2p_i\delta \geq f_{\lambda-1}(y_t) + 1 + \delta$. The last inequality is true since $\delta \leq 1$ and $p_i \leq 1$. Thus, $vs(\{x_t, x_b\})$ increases by at most $(p_{x_b} + p_{x_t})2p_i\delta$ while $vs(\{y_t, y_b\})$ decreases by $(p_{y_t} + p_{y_b})2(1 - p_i)\delta = (p_{x_b} + p_{x_t})2p_i\delta$, ensuring that violation score along the vertical dimension does not increase. \square

4.2.1 Proof of Lemma 4.3

Using the arguments very similar to [JR11] as given below, we can get the following sequence of inequalities

$$\begin{aligned} Dist(f_{i-1}, f_i) &= Dist(f_{i-1}, A_i(f_{i-1})) \leq \sum_{(x,y) \in V_i(f_{i-1})} (p_x + p_y) \\ &\leq \frac{1}{\delta} VS^i(f_{i-1}) \leq \frac{1}{\delta} VS^i(f) + 2(d - i)\delta \leq \frac{1}{\delta} \sum_{(x,y) \in V^i(f)} (p_x + p_y) \cdot ImD(f) + 2(d - i)\delta \end{aligned}$$

Here functions $\{f_i\}_{i=0}^d$ are defined in the same way as [JR11]. The first inequality holds because A_i modifies f only at the endpoints points x and y of violated edge (x, y) along dimension i , thus paying $p_x + p_y$. The second and fourth inequalities follow from Equation (3) and the third inequality holds because of Lemma 4.9. Therefore, by triangle inequality, we have

$$\begin{aligned} Dist(f, f_d) &\leq \sum_{i \in [d]} Dist(f_{i-1}, f_i) \leq \sum_{i \in [d]} \left(\sum_{(x,y) \in V^i(f)} (p_x + p_y) \cdot \frac{ImD(f)}{\delta} \right) + 2(d - i)\delta \\ &\leq \left(\sum_{(x,y) \in V(f)} (p_x + p_y) \cdot \frac{ImD(f)}{\delta} \right) + d^2\delta \end{aligned}$$

For a function which is ϵ -far from Lipschitz, we have $Dist(f, f_d) \geq \epsilon$. Therefore, from the above inequality, we have

$$\sum_{(x,y) \in V(f)} \frac{(p_x + p_y)}{d} \geq \frac{\delta(\epsilon - d^2\delta)}{d \cdot ImD(f)}$$

5 Instantiation of privacy tester using Lipschitz testers

In this section, we instantiate the privacy tester of Section 3 with both known Lipschitz testers as well as the Lipschitz tester developed in this work. This is presented in the table below. The third column gives the “approximation factor” as defined in Definition 3.2 for the various testers. The final column gives the privacy tester parameters that each of the tester achieves. The last row gives the result of Lipschitz tester (Section 4) developed in this work.

Reference	Functions	Approximation factor (θ)	Distribution	Tester running time	Privacy tester
[JR11]	$\{0, 1\}^d \rightarrow \mathbb{R}$	$1 + \delta$	Uniform	$O\left(\frac{d \cdot \text{ImD}(f)}{\epsilon \delta}\right)$	$(1 + \delta, \alpha, \gamma, \beta)$
[AJMR12b]	$\{1, \dots, n\}^d \rightarrow \mathbb{R}$	$1 + \delta$	Uniform	$\tilde{O}\left(\frac{d \min\{\text{ImD}(f), nd\}}{\delta \epsilon}\right)$	$(1 + \delta, \alpha, \gamma, \beta)$
[CS12]	$\{0, 1\}^d \rightarrow \mathbb{R}$	1	Uniform	$O\left(\frac{d}{\epsilon}\right)$	$(1, \alpha, \gamma, \beta)$
This work	$\{0, 1\}^d \rightarrow \mathbb{R}$	$1 + \delta$	Product	$O\left(\frac{d \cdot \text{ImD}(f)}{(\epsilon - d^2 \delta) \delta}\right)$	$(1 + \delta, \alpha, \gamma, \beta)$

6 Discussions and Open Problems

In this section we discuss about some of the interesting implications of our current work and some of the new avenues it opens up. Also we state some of the open problems that remains unresolved in our work.

Privacy: In this work, we took the first step towards designing efficient testing algorithm for statistical data privacy. Our work indicates that it is indeed possible to design efficient testing algorithms for some existing notions of statistical data privacy (e.g., generalized differential privacy). It is important that the current paper should be treated as an initial study of the problem and in no way should be interpreted conclusive. It is interesting to explore other rigorous notions of data privacy, their applications and design testers for them.

In this paper, we test for generalized differential privacy, which is a relaxation of differential privacy. It remains an open problem to design a privacy tester for exact differential privacy. The problem seems to be challenging because of the fact that if we want to design an efficient tester, then usually the utility guarantees for the tester allow it to fail with some probability. Now, differential privacy being a worst case notion, it is not clear how to incorporate the failure property of the tester and yet make precise claims about differential privacy.

In the current work, we have designed privacy testers for algorithms where the domain of the data sets are either hypercube or hypergrid. A natural question that arises is that if we can extend the current results to design privacy testers when the data sets are drawn from continuous domain, unlike hypercube or hypergrid.

Lipschitz Testing: This work presents the first Lipschitz property tester for the setting where the domain points are sampled from a distribution that is not uniform. Because of possible applications to statistical data privacy, this work has motivated the design of such Lipschitz testers for other domains, e.g. hypergrid. Also, this paper mainly shows the tester for the product distribution over the hypercube domain, but it still remains open to design testers for other distributions that may be correlated in some way (e.g., pairwise correlation).

Acknowledgements: We would like to thank Sofya Raskhodnikova and Adam Smith for various suggestions and comments during the course of this project.

References

- [AC06] Nir Ailon and Bernard Chazelle. Information theory in property testing and monotonicity testing in higher dimension. *Inf. Comput.*, 204(11):1704–1717, 2006.
- [AJMR12a] Pranjal Awasthi, Madhav Jha, Marco Molinaro, and Sofya Raskhodnikova. Limitations of local filters of lipschitz and monotone functions. In Gupta et al. [GJRS12], pages 387–398.

- [AJMR12b] Pranjal Awasthi, Madhav Jha, Marco Molinaro, and Sofya Raskhodnikova. Testing lipschitz functions on hypergrid domains. In Gupta et al. [GJRS12], pages 387–398.
- [BBG⁺11] Raghav Bhaskar, Abhishek Bhowmick, Vipul Goyal, Srivatsan Laxman, and Abhradeep Thakurta. Noiseless database privacy. In Dong Hoon Lee and Xiaoyun Wang, editors, *ASIACRYPT*, volume 7073 of *Lecture Notes in Computer Science*, pages 215–232. Springer, 2011.
- [BD12] Abhishek Bhowmick and Cynthia Dwork. Natural differential privacy. In *Personal communication*, 2012.
- [CKN⁺11] Joseph A. Calandrino, Ann Kilzer, Arvind Narayanan, Edward W. Felten, and Vitaly Shmatikov. ”you might also like: ” privacy risks of collaborative filtering. In *IEEE Symposium on Security and Privacy*, 2011.
- [CS12] Deeparnab Chakrabarty and C. Seshadhri. Optimal bounds for monotonicity and lipschitz testing over the hypercube. *CoRR*, abs/1204.0849, 2012.
- [DGL⁺99] Yevgeniy Dodis, Oded Goldreich, Eric Lehman, Sofya Raskhodnikova, Dana Ron, and Alex Samorodnitsky. Improved testing algorithms for monotonicity. In *RANDOM*, pages 97–108, 1999.
- [DKMN06] Cynthia Dwork, Krishnamurthy Kenthapadi, Frank McSherry, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, pages 486–503. Springer, 2006.
- [DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.
- [Dwo06] Cynthia Dwork. Differential privacy. In *ICALP*, 2006.
- [Dwo08] Cynthia Dwork. Differential privacy: A survey of results. In *TAMC*, pages 1–19. Springer, 2008.
- [Dwo09] Cynthia Dwork. The differential privacy frontier. In *TCC*, pages 496–502. Springer, 2009.
- [GGL⁺00] Oded Goldreich, Shafi Goldwasser, Eric Lehman, Dana Ron, and Alex Samorodnitsky. Testing monotonicity. *Combinatorica*, 20(3):301–337, 2000.
- [GGR98a] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998.
- [GGR98b] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998.
- [GJRS12] Anupam Gupta, Klaus Jansen, José D. P. Rolim, and Rocco A. Servedio, editors. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings*, volume 7408 of *Lecture Notes in Computer Science*. Springer, 2012.
- [GKS08] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, pages 265–273, 2008.
- [GS09] Dana Glasner and Rocco A. Servedio. Distribution-free testing lower bound for basic boolean functions. *Theory of Computing*, 5(1):191–216, 2009.
- [HK07] Shirley Halevy and Eyal Kushilevitz. Distribution-free property-testing. *SIAM J. Comput.*, 37(4):1107–1138, 2007.
- [JR11] Madhav Jha and Sofya Raskhodnikova. Testing and reconstruction of lipschitz functions with applications to data privacy. In Rafail Ostrovsky, editor, *FOCS*, pages 433–442. IEEE, 2011.
- [Kor10] Aleksandra Korolova. Privacy violations using microtargeted ads: A case study. In *ICDMW*, 2010.
- [McS09] Frank D. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *SIGMOD*, 2009.
- [MGKV06] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. 1-diversity: Privacy beyond k-anonymity. In *ICDE*, page 24, 2006.
- [MTS⁺12] Prashanth Mohan, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David Culler. Gupt: privacy preserving data analysis made easy. In *SIGMOD*, 2012.
- [NRS07] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, 2007.

- [RP10] Jason Reed and Benjamin C. Pierce. Distance makes the types grow stronger: a calculus for differential privacy. In *ICFP*, 2010.
- [RS96a] Ronitt Rubinfeld and Madhu Sudan. Robust characterization of polynomials with applications to program testing. *SIAM J. Comput.*, 25(2):252–271, 1996.
- [RS96b] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM J. Comput.*, 25(2):252–271, 1996.
- [RSK⁺10] Indrajit Roy, Srinath T. V. Setty, Ann Kilzer, Vitaly Shmatikov, and Emmett Witchel. Airavat: Security and privacy for mapreduce. In *NSDI*, 2010.
- [Swe02] Latanya Sweeney. k -anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.